BMJ Open

Which factors may determine the necessary and feasible type of effectiveness evidence? A mixed methods approach to develop an instrument to help coverage decision-makers

Saskia de Groot,[1] Adriana J Rijnsburger,[1] Matthijs M Versteegh,[1] Juanita M Heymans,[2] Sarah Kleijnen,[2] W Ken Redekop,[1] Ilse M Verstijnen[2]

To cite: de Groot S, Rijnsburger AJ, Versteegh MM, *et al*. Which factors may determine the necessary and feasible type of effectiveness evidence? A mixed methods approach to develop an instrument to help coverage decision-makers. *BMJ Open* 2015;**5**:e007241. doi:10.1136/bmjopen-2014-007241

▸ Prepublication history for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2014-007241).

Received 18 November 2014
Revised 2 April 2015
Accepted 12 April 2015

[1]Department of Health Policy and Management, Institute for Medical Technology Assessment, Erasmus University Rotterdam, Rotterdam, The Netherlands
[2]Dutch National Health Care Institute (ZIN) (formerly named CVZ), Diemen, The Netherlands

Correspondence to
Saskia de Groot;
s.degroot@bmg.eur.nl

ABSTRACT

Objectives: Reimbursement decisions require evidence of effectiveness and, in general, a blinded randomised controlled trial (RCT) is the preferred study design to provide it. However, there are situations where a cohort study, or even patient series, can be deemed acceptable. The aim of this study was to develop an instrument that first examines which study characteristics of a blinded RCT are necessary, and then, if particular characteristics are considered necessary, examines whether these characteristics are feasible.

Design: We retrospectively studied 22 interventions from 20 reimbursement reports concerning medical specialist care made by the Dutch National Health Care Institute (ZIN) to identify any factors that influenced the necessity and feasibility of blinded RCTs, and their constituent study characteristics, that is, blinding, randomisation and a control group. A literature review was performed to identify additional factors. Additional expertise was included by interviewing eight experts in epidemiology, medicine and ethics. The resulting instrument was called the FIT instrument (Feasible Information Trajectory), and was prospectively validated using three consecutive reimbursement reports.

Results: (Blinded) RCT evidence was lacking in 5 of 11 positive reimbursement decisions and 3 of 11 negative decisions. In the reimbursement reports, we found no empirical evidence supporting situations where a blinded RCT is unnecessary. The literature also revealed few arguments against the necessity of a blinded RCT. In contrast, many factors influencing the feasibility of randomisation, a control group and blinding, were found in the reimbursement reports and the literature; for example, when a patient population is too small or when an intervention is common practice, randomisation will be hindered.

Conclusions: Policy regarding the necessity and feasibility of different types of evidence of effectiveness would benefit from systematic guidance. The FIT instrument has the potential to support transparent, reproducible and well-founded decisions on appropriate evidence of effectiveness in medical specialist care.

Strengths and limitations of this study

- In this study, multiple sources were used, including 20 reimbursement reports made by the Dutch National Health Care Institute, along with literature and expert opinion.
- Since most items used to examine which study characteristics of a blinded randomised controlled trial remain necessary and feasible are based on general epidemiological principles, results of this study might also be useful for reimbursement agencies in countries other than the Netherlands.
- Not all possible study characteristics are taken into account; the study was limited to the necessity and feasibility of randomisation, a control group and blinding.
- The instrument's completeness requires further testing, as does its impact on the decision-making process in terms of efficiency, reliability (such as inter/intraobserver reliability) and additional forms of validity.

INTRODUCTION

Reimbursement decisions require evidence of effectiveness. To demonstrate effectiveness, a blinded randomised controlled trial (RCT) is, in general, the preferred study design. The central reimbursement authority in the Netherlands, the Dutch National Health Care Institute (ZIN), formerly named Health

Care Insurance Board (CVZ), applies the principles of evidence-based medicine (EBM) to determine whether care is effective.[1] EBM is "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients".[2] It integrates the best available external evidence with individual clinical expertise and patient preferences. In 2015, ZIN will update its framework for decision-making, and formally integrate the 'Grading of Recommendations Assessment, Development and Evaluation' (GRADE) system to assess and grade available evidence.[3]

Although blinded RCTs are often preferred, they are not always available, and reimbursement decisions have to be taken despite suboptimal evidence. A systematic review by Fischer[4] showed that, indeed, the presence of suboptimal evidence plays an important role in coverage decision-making. Whereas evidence from RCTs is absolutely necessary to overcome problems such as confounding by indication, it has been stressed that, in some situations, an RCT may be unnecessary, inappropriate, impossible or inadequate.[5] For example, in situations where successful interventions for otherwise fatal conditions become available, an RCT is not required.[2] Observational studies may provide valuable information, in particular on rare or long-term harms,[6] but also on the benefits of an intervention.

The assessment of the quality of available external evidence has received much scientific attention resulting in elaborate rating methods such as the GRADE system.[7] An assessment of the necessity and feasibility of attaining certain evidence requirements has, by contrast, received less attention. The aim of this study was to develop an instrument that first examines which study characteristics of a blinded RCT are necessary, and then, if particular characteristics are considered necessary, examines whether these characteristics are feasible.[8] Together, the necessary and feasible characteristics define the optimal study design. A blinded RCT is, by consequence, necessary and feasible when all three characteristics (blinding, randomisation and a control group) are considered necessary and feasible. If they are not, a different study design may be considered optimal.

We focused on evidence of effectiveness for non-pharmaceutical, therapeutic medical specialist care as a starting point. Although the instrument is based on this type of care, it may be applicable to other types of interventions.

METHODS

Three information sources were used to identify factors influencing the necessity and feasibility of different types of evidence of effectiveness. These sources are specified below.

Review of reimbursement reports

All medical specialist care reports published by ZIN between 1 January 2007 and 31 December 2010 were collected. Reports were available on all reimbursement decisions. 1 January 2007 was chosen as a starting point, since from this date onwards, ZIN officially applied the principles of EBM to determine whether care is effective. As data extraction was performed in 2011, 2010 was the last year for which complete reports were available. A stratified sample was implemented in order to prevent ending up with a sample only including reimbursement reports for which the necessity and feasibility of blinded RCTs was evident. The reimbursement reports were therefore classified into three groups based on their level of complexity, namely, simple (few necessity or feasibility issues), intermediate (moderate necessity or feasibility issues) or complex (complex necessity or feasibility issues), by the three researchers from ZIN (JH, SK, IV). Subsequently, a random sample of each group was drawn (simple N=5, intermediate N=7 and complex N=8). The resulting sample concerned 22 medical specialist care interventions (two reports evaluated two interventions). Then, data from these reports were extracted in a uniform matter on the following items: the intervention, the patient group, the reimbursement decision (negative or positive), the available evidence of effectiveness, and any arguments that influenced the necessity and feasibility of different types of evidence of effectiveness. A negative or positive reimbursement decision implies whether the intervention was considered suitable for reimbursement from the basic health insurance package in the Netherlands. These decisions did not always rely on the assessment of effectiveness only, but may take additional reimbursement criteria into account.

Literature review

A literature review in Medline (1996 to March 2011) was performed to identify publications that discussed factors influencing the necessity and feasibility of different types of evidence. The Ovid search interface was used. The search strategy included the following keyword combinations: level? or degree? or criteri$ or hierarch$ or require$ or assess$ or standard$ and evidence-based practice. The resulting publications were screened systematically. Selection on relevance was carried out by two researchers (SdG, AR). First, titles and abstracts were screened; publications that might discuss any factors influencing the necessity and feasibility of different types of evidence of effectiveness were included. The full texts of the selected publications were then examined; publications that did not discuss any factors influencing the necessity and feasibility of different types of evidence of effectiveness were excluded, as were publications not in English and publications not available as full text. The reference lists of relevant publications were screened by hand to identify any additional publications for inclusion. Special attention was given to those factors affecting the constituent study characteristics of blinded RCTs; namely, randomisation, a control group and blinding, because there are factors that may influence one of

the characteristics but not the other. Each of these characteristics by itself is meant to reduce bias, hence the more these characteristics are preserved (up to a full-blown blinded RCT) the less biased the research outcomes will be. Obviously, less biased outcomes form better grounds for decision-making.

Interviews

A draft instrument was drawn based on the factors influencing the necessity and feasibility of different types of evidence of effectiveness found in the reimbursement reports and in the literature. Subsequently, additional expertise was included by interviewing eight experts in epidemiology, medicine and ethics. All participants were informed about the purpose and content of the interview and agreed to participate by email. Interviews were semi-structured, that is, questions were predefined, but interviewees were allowed to raise other relevant issues not covered by the interview schedule.[9] Seven of eight interviews were performed face-to-face, and one was by telephone. The interviews were intended to elicit views and opinions from respondents on the factors included in the draft instrument, and to identify additional factors.

Figure 1 shows the steps of the research process, including its aims.

Definitions of necessity and feasibility

The resulting instrument was called the FIT instrument (Feasible Information Trajectory). The factors included were categorised in two groups: The first group deals with the necessity of randomisation, a control group and blinding, relating to situations in which one or more of these three characteristic are not required. The second group deals with the feasibility of randomisation, a control group and blinding, and was subdivided into two groups (2A and 2B). Group 2A refers to factors that, stand alone, are sufficiently strong to deviate from randomisation, a control group and/or blinding. Group 2B refers to factors that, by themselves, are insufficiently strong, but may jointly provide a compelling case to do so.

Prospective validation

As a last step, the FIT instrument was prospectively validated by four ZIN decision-makers who were not involved in the study. They each applied the instrument on one of three consecutive reimbursement reports regarding medical specialist care which included immunotherapy for high-risk neuroblastoma, zygomatic implants for the atrophic edentulous maxilla and endovascular treatment of complex aortic aneurysms. The instrument's face validity was discussed in a joint meeting with the project team. The main question in this respect was whether the instrument contained all facets to identify the necessity and feasibility of the constituent study characteristics of blinded RCTs. Therefore, the relevance and clarity of the factors included in the instrument, as well as the completeness and user-friendliness of the instrument, were discussed. Instructions about the validation were sent to the decision-makers beforehand. These instructions explained the aim of the validation, and pointed out particular questions to the decision-makers, such as 'Do all questions apply?', 'Do you miss any questions?' and 'Are the questions clearly stated?'. Suggestions for improvement of the instrument were applied if everyone (the decision-makers and the project team) agreed.

The final FIT instrument is programmed in Excel and requires the user to answer questions regarding the factors that influence the necessity and feasibility of randomisation, a control group and blinding.

Since this article does not contain any personal medical information about an identifiable living individual, ethics approval was not required.

RESULTS

Eleven of 22 reimbursement decisions were negative, that is, the intervention was not considered suitable for reimbursement from the basic health insurance package in the Netherlands, and the remaining 11 decisions were positive. In 5 of 11 positive reimbursement decisions, evidence from (blinded) RCTs was lacking. Arguments that contributed to the acceptance of the available evidence in these cases are presented in table 1. The arguments listed in the table are the only arguments explicitly mentioned in the reimbursement reports.

Of the 11 negative reimbursement decisions, evidence from (blinded) RCTs was lacking in three of them. Besides suboptimal evidence of effectiveness, these three decisions revealed arguments against the acceptance of

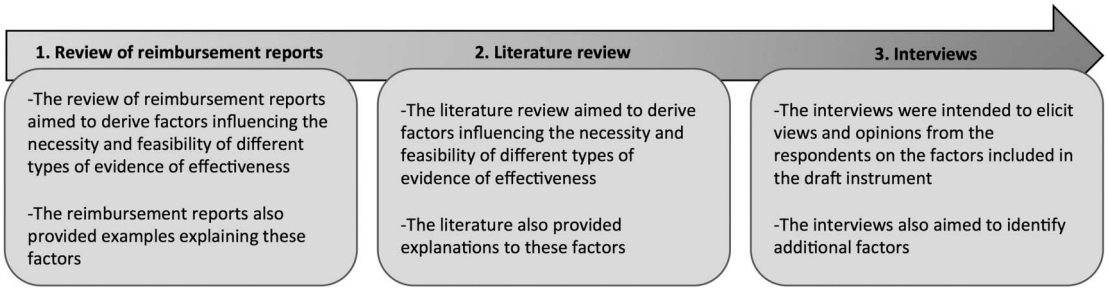

Figure 1 Steps of the research process including its aims.

the available evidence, such as poor quality of the available evidence and limited evidence on long-term safety (table 2). The arguments listed in the table are the only arguments explicitly mentioned in the reimbursement reports. Apparently, there were no decisive arguments to accept suboptimal evidence.

In some reimbursement decisions, additional reimbursement criteria besides effectiveness were taken into account. These criteria could have influenced the reimbursement decision, but are not included in table 2.

Evidence from (blinded) RCTs was available in 14 of 22 reimbursement decisions (six positive and eight negative). Negative decisions were mostly motivated by the insufficient quality of the RCT, a lack of long-term evidence, and absence of consensus within the literature or the profession on the effectiveness and safety of the intervention. Again, besides effectiveness, other reimbursement criteria could have played a role.

The literature search in Medline identified 98 publications. Seventy-eight publications were discarded based on their title or abstract. Additionally, three publications were excluded, because no full text was available. Another three publications were excluded, because they were not written in English. The full texts of the

Table 1 Positive reimbursement decisions in reports with suboptimal evidence of effectiveness; arguments that advocated for the acceptance of the available suboptimal evidence

Intervention	Arguments that advocated for the acceptance of the available suboptimal evidence‡
I. Metal on metal (MoM) resurfacing arthroplasty of the hip for patients with primary or secondary osteoarthritis*[10]	▶ **Technique has already been used** for approximately 10 years ▶ **Patients will probably not participate in a randomised study** given the available data ▶ **Availability of lower level evidence of effectiveness** (comparative studies, short to medium term)
II. Preimplantation genetic diagnosis (PGD) to predict β-thalassaemia in second child with human leucocyte antigen (HLA) typing during PGD for stem cell transplantation[11]	▶ **PGD as a technique has been proven effective** in predicting other diseases ▶ **It is unfeasible to perform comparative studies** on the use of PGD to predict β-thalassaemia ▶ **It is unfeasible to perform comparative studies** of HLA typing
III. Potassium-titanyl-phosphate (KTP) laser treatment for patients with lower urinary tract symptoms (LUTS) caused by benign prostatic hyperplasia†[12]	▶ **(Double)-blinding is unfeasible** in comparing KTP laser treatment with transurethral resection of the prostate (TURP) for patients with a mildly enlarged prostate (surgical intervention) ▶ **Technique is already frequently used** ▶ **Patients will probably not participate in a randomised study** ▶ **No alternative treatment** if drugs provide no adequate relief (for high-risk patients or patients being treated with anticoagulants) ▶ **Availability of lower level evidence of effectiveness and safety** (only non-comparative studies for high-risk patients)
IV. Transcatheter pulmonary valve implantation (TPVI) for patients with an abnormal pulmonary valve due to a congenital heart defect[13]	▶ **(Double)-blinding is unfeasible** in comparing TPVI with surgical pulmonary valve replacement ▶ **Rare condition;** only 30 patients are eligible for percutaneous pulmonary valve implantation in the Netherlands per year ▶ **Safety concerns seem limited** regarding percutaneous pulmonary valve implantation, particularly in comparison to open cardiac surgery ▶ **Availability of lower level evidence of effectiveness and safety,** case series showed good short-term success rates
V. Proton therapy for patients with intraocular tumours, chordomas and chondrosarcomas, and paediatric tumours[14]	▶ **Side effects are rare and mostly happen (many) years later,** and the purpose of the therapy, in particular, is to reduce or prevent late side effects (instead of to prove effectiveness) ▶ **Rare conditions** ▶ **Technique is already frequently used** and there is international consensus between radiation therapists and oncologists ▶ **Availability of lower level evidence of effectiveness** (mainly case series)

*This intervention complied with medical science and medical practice, according to the Dutch National Health Care Institute (ZIN), for patients with primary or secondary osteoarthritis who failed on conservative treatment, or patients younger than 65 years of age with a sufficient level of activity. However, in January 2012, this advice was changed in line with the advice of the Netherlands Orthopaedic Association, based on new national and international published experiences, to not place large-head MoM hip implants and MoM resurfacing implants.
†This intervention complied with medical science and medical practice, according to ZIN, for patients with a mildly enlarged prostate, high-risk patients or patients being treated with anticoagulants.
‡Essential arguments are shown in bold.

Table 2 Negative reimbursement decisions in reports with suboptimal evidence of effectiveness; arguments that advocated against the acceptance of the available evidence

Intervention	Arguments that advocated against the acceptance of the available evidence†
VI. Transcatheter aortic valve implantation (TAVI) for patients with aortic valve stenosis*[13]	▶ **Substantial intervention-related mortality**, probably (partly) related to a priori risk profile of study population ▶ **No comparative studies available** and the quality of the available evidence is poor (ie, limited duration and limited size) ▶ **Randomised controlled trial (RCT) is on-going**
VII. Psychoanalysis[15]	▶ **Insufficient studies (qualitatively adequate) on the effectiveness available** ▶ Existence of several comparative effectiveness studies on long-term psychoanalytic psychotherapy makes it likely that such studies are also possible for psychoanalysis
VIII. Breast augmentation with autologous lipofilling[16]	▶ **Insufficient studies on the effectiveness available** (majority are case reports and non-comparative studies) ▶ **No consensus** on breast augmentation with autologous lipofilling (unclear whether possible microcalcifications influence the assessment of mammograms)

*In October 2011, this advice was changed. In the opinion of the Dutch National Health Care Institute (ZIN), TAVI is care that complies with established medical science and medical practice for insured persons with severe stenosis of the aortic valve and for whom the surgical risks are unacceptably high. TAVI belongs in the insured basic package for these insured persons.
†Essential arguments are shown in bold.

remaining 14 publications were examined. Six publications were excluded since they did not provide any factors influencing the necessity and feasibility of different types of evidence of effectiveness. The resulting eight publications revealed overlapping, but also showed factors in addition to those found in the reimbursement reports.

Table 3 comprises the final FIT instrument, and summarises factors influencing the necessity and feasibility of the constituent study characteristics of blinded RCTs, derived from the reimbursement reports, the literature review and the interviews.

When is randomisation, a control group or blinding necessary?

In RCTs, patients are allocated randomly to one of the interventions being studied, to minimise imbalances in confounding variables between experimental and control groups.[17–20] The merits of a random allocation procedure are particularly important if, for example, confounding by indication is likely. This happens when the indication for an intervention is related to the health outcome and, as a result, the effect estimate is distorted because it is mixed with the effect of a confounding variable. A systematic review and meta-analysis of drug-eluting stents versus bare metal stents showed that the rates of death and myocardial infarction were found to be significantly reduced with the use of drug-eluting stents in the observational studies with an attenuated effect in the RCTs.[21] One of the explanations for this reduction, as mentioned by the authors, was the non-randomised choice of either drug-eluting stents or bare metal stents. The decision to use drug-eluting stents may be based on unmeasured patient characteristics and may importantly affect subsequent treatment decisions, including medication use.

A control group for the intervention of interest can involve placebo, active treatment, no treatment, a different dose or regimen of the study treatment or external or historical controls.[17] A control group is particularly needed if the disease has a favourable natural history, for example, when complaints naturally diminish over time. Uncontrolled or poorly controlled studies tend to overestimate the treatment effect, because they do not take into account this favourable natural history of the disease.[22]

Blinding of study participants, health providers and investigators reduces the risk that knowledge of which intervention was received, rather than the intervention itself, affects outcomes.[23] Blinding of outcome assessors combined with a control group is particularly needed if the primary outcome measure is subjective, such as quality of life, instead of objective, such as recurrence from disease or even death.[17 20 23] Blinding of patients and physicians prevents them, for example, from differential drop-out or differential administration of cointerventions, thereby eliminating possible effects of differential behaviours across intervention groups.[17 20 23]

There may, however, be conditions where randomisation and a control group are not specifically needed, for example, in case of a 'dramatic' or 'immediate' effect, since it is less likely that the effect will be explained by (unknown) confounders only.[17 19 20] Insulin supplementation in diabetic patients is an example in which the effect rapidly follows the intervention. A 'dramatic' or 'immediate' effect in combination with a plausible relationship between the pathophysiology of the condition

Table 3 Final FIT instrument summarising factors influencing the necessity and feasibility of randomisation, a control group and blinding

	Randomisation?	Control group?	Blinding of outcome assessors?	Blinding of patients and physicians?	Reference to reimbursement reports	Reference to the literature
1. When is randomisation, a control group or blinding necessary?						
▶ Confounding by indication	+	+				17–20
▶ Natural decrease of symptoms over time		+			35	22
▶ Subjective outcome measures		+	+			17 20 23
▶ Differential behaviours across intervention groups				+		17 20 23
▶ Dramatic/immediate effects	–	–				17 19 20
2A. When is randomisation, a control group or blinding hard to achieve?						
▶ Lack of equipoise, that is, consensus about the preferred treatment	–				Report I, III[10 12]	25 26
▶ Outcomes occur in the distant future	–			–		17 18 22 27
▶ Small adaptation of an intervention that has already been proven effective and reimbursed	–	–				25
▶ Extension of the indication area of a procedure that has already been proven effective and reimbursed	–	–			Report II[11]	
2B. Which factors hinder the feasibility of randomisation, a control group or blinding?						
▶ Small patient population	–				Report IV, V[13 14]	17 25
▶ Poor prognosis/no alternative treatment	–	–			Report III[12]	29
▶ Intervention is common practice	–				Report I, III, V[10 12 14]	25 29
▶ Urgency of the intervention	–			–		25
▶ Complexity of the intervention	–			–	33	27
▶ Availability of good quality low level evidence of effectiveness	–	–	–	–	Report I, III, IV, V[10 12–14]	

A plus sign indicates that randomisation, a control group or blinding is necessary. A minus sign indicates that randomisation, a control group or blinding is unnecessary, is hard to achieve or is hindered.
FIT, Feasible Information Trajectory.

and the mechanism of the intervention may legitimise suboptimal evidence. Besides plausibility, additional factors contribute to our belief that the relationship between an intervention and its effect is indeed causal, such as consistency (across research settings), temporality (effect follows treatment) and biological gradient (dose–response relation).[24]

When is randomisation, a control group or blinding hard to achieve?

First, randomisation may be hard to achieve if equipoise is lacking,[25] that is, if there is consensus about the preferred treatment. In this case, patients and clinicians would both be unwilling to participate in a randomised study. Furthermore, ethicists claim that it would be unethical to randomise patients.[26]

Second, if outcomes occur in the distant future and intermediate outcome measures are absent, an extended follow-up is needed, which complicates randomisation.[17 18 22 27] Also, blinding of patients and physicians will be problematic, since eventually, for example, physicians may need to know what treatment the patient has received in deciding on future treatment.

In addition, in case of a small adaptation of an intervention that has already been proven effective and reimbursed, randomisation and a control group may also be hard to achieve. Surgical techniques generally progress through a number of successive, small modifications.[25]

Strictly speaking, evidence from RCTs is still required to demonstrate effectiveness unless there is a minimal chance that the treatment effect will be poorer than the effect previously demonstrated. If evidence from RCTs is required, there are reasons why randomisation and a control group may be hard to achieve. First, if there is an imbalance between the amount of effort needed to provide this evidence and the degree of reduction in uncertainty about the effectiveness of the intervention, conducting an RCT will not be very rational. Second, if RCTs were needed to evaluate each small modification, progress would be slowed.[25] Third, excluding half of the population from the new intervention in order to reduce a small degree of uncertainty about the treatment effect might not be ethical. If the reimbursement assessment involves an extension of the indication area of a procedure that has already been proven effective and reimbursed for a smaller indication, this may also legitimise the use of suboptimal evidence.[11]

Lastly, characteristics of the intervention could make blinding of the outcome assessors as well as blinding of patients and physicians impossible. For example, with most surgical interventions it is impossible to blind the surgeon. Also, blinding of patients may be impossible, as was seen in the reimbursement report regarding sacral neurostimulation in patients with faecal incontinence. Since this device sends mild electrical pulses to the sacral nerves that control the bowel and rectum, patients will feel the treatment they receive, making blinding impossible.[28] When blinding of patients and physicians is impossible, it still might be possible to blind the outcome assessors, especially if the intervention is not visible.

Which factors hinder the feasibility of randomisation, a control group or blinding?

Besides factors that make randomisation, a control group or blinding unnecessary or hard to achieve, there may be several characteristics of the patient population and the intervention that hinder the feasibility of randomisation, a control group or blinding. Factors that hinder the feasibility of randomisation, a control group or blinding may by themselves be insufficiently strong, but may jointly argue for deviating from randomisation, a control group or blinding.

First, a small patient population limits the feasibility of an RCT with sufficient power to detect a real effect as statistically significant.[17 25] This was suggested in the reimbursement report on the transcatheter pulmonary valve implantation and in the reimbursement report on proton therapy for patients with intraocular tumours, chordomas and chondrosarcomas and paediatric tumours.[13 14] While international trials are performed to obtain sufficient power in orphan drug research, performing international trials for non-pharmaceutical medical specialist care may be complicated, especially when treatments, such as surgical procedures, vary across borders.

Second, a poor prognosis in combination with withholding the only possibly effective treatment could hinder the feasibility of randomisation and a control group, and could therefore legitimatise the acceptance of suboptimal evidence.[29] If treatment is urgent and patients experience a significant burden of the disease while the new intervention provides the only chance on improvement, one could argue that it would be unethical to randomise patients to either doing nothing or the intervention. This was proposed in appraising the effectiveness of KTP laser treatment for high-risk patients with benign prostatic hyperplasia, since drugs do not always provide adequate relief for these patients.[12]

Third, an intervention may be common practice without the availability of evidence from blinded RCTs.[25 29] These interventions are frequently applied in clinical practice and sometimes even included in clinical guidelines. Performing an RCT would be considered as 'outdated', and randomisation would therefore be hindered. Heart, liver, kidney and lung transplantations are examples of interventions that are not validated with RCTs.[30] This argument was also mentioned in the reimbursement reports on metal on metal resurfacing arthroplasty, treatment with KTP laser and proton therapy.[10 12 14]

Fourth, the urgency of an intervention, that is, time urgency, could hinder randomisation and obtaining informed consent.[25] Fifth, the feasibility of an RCT may be limited because of the complexity of interventions.[27] Complex interventions are generally described as interventions that contain several interacting components, but according to Craig *et al*,[31] complex interventions have some additional characteristics including the "number and difficulty of behaviours required by those delivering or receiving the intervention, the number of groups or organisational levels targeted by the intervention, the number and variability of outcomes and the degree of flexibility or tailoring of the intervention permitted". The evaluation of a specialist stroke unit is an example, since such an evaluation would have to consider the expertise of various health professionals as well as investigations, drugs, treatment guidelines, and arrangements for discharge and follow-up. Furthermore, the organisation, management and skill mix of stroke units differ.[32] In the reimbursement report on rheopheresis, complexity was put forward to argue that blinding became unfeasible.[33]

Lastly, the availability of good quality low level evidence may limit the feasibility of randomisation, a control group and blinding, since this influences both patient and clinical equipoise. Moreover, if this evidence is methodologically strong, and shows large and consistent results, we may be confident about the results,[34] and the necessity of randomisation, a control group and blinding may also be limited.

Prospective validation

During the prospective validation, four ZIN decision-makers applied the FIT instrument to assess three new

interventions. In assessing immunotherapy for high-risk neuroblastoma, the FIT instrument revealed that blinding of outcome assessors, patients and physicians would be hard to achieve as a result of the characteristics of the intervention, including its intensity. Although some factors that hinder randomisation or a control group were identified, such as a small patient population and a poor prognosis, randomisation and a control group were still considered necessary and feasible. Blinding of outcome assessors, patients and physicians also appeared hard to achieve as a result of the characteristics of the second intervention (zygomatic implants for the atrophic edentulous maxilla). Although the FIT instrument revealed some factors that hinder randomisation and/or a control group, such as a small patient population and complexity of the intervention, randomisation and a control group were still considered necessary and feasible. In assessing endovascular treatment of complex aortic aneurysms, a small patient population and a poor prognosis, combined with the absence of an alternative treatment for some of the patients, were identified by the FIT instrument as factors that hinder randomisation and a control group.

In general, the ZIN decision-makers recognised the factors included in the instrument and they all found the instrument useful in deciding what can be considered the appropriate evidence of effectiveness. Based on discussions with the decision-makers, two factors hindering the feasibility of randomisation, a control group or blinding were excluded: lack of funding and lack of a good research infrastructure (sufficient expertise, facilities and materials). The reason for these exclusions was that these factors have an organisational nature, whereas all other factors relate to characteristics of the patient population or the intervention. None of the decision-makers contributed additional factors, the instrument therefore appeared complete.

DISCUSSION

As reimbursement agencies are under pressure to make rapid and well-founded decisions on reimbursement, there is a need for systematic guidance on what can be considered the appropriate evidence of effectiveness. We therefore developed an instrument that first examines which study characteristics of a blinded RCT are necessary, and then, if particular characteristics are considered necessary, examines whether these characteristics are feasible. As shown in table 3, in the reimbursement reports we found no empirical evidence supporting situations where a blinded RCT is unnecessary. The literature revealed only one situation in which a blinded RCT was not required, where, in case of a 'dramatic' or 'immediate' effect, randomisation and a control group become unnecessary. In contrast, many factors influencing the feasibility of randomisation, a control group and blinding were identified in both the literature and the reimbursement reports. One could therefore argue that a blinded RCT is almost always necessary, but not always feasible.

Recently, Van Loon *et al*[36] considered five arguments that limit the feasibility of RCTs in the evaluation of novel radiotherapy technologies, namely, rare indications, narrow inclusion criteria, end points that require data for late toxic effects or second malignant disease, limited funding and a strong belief in effectiveness of the novel technique. Additionally, the authors proposed guidelines for prospective comparative cohort studies when RCTs are not feasible. The FIT instrument reveals which study characteristics of a blinded RCT are necessary and feasible resulting in a richer research palette of study designs that go beyond prospective comparative cohort studies. For example, differentiation between controlled observational studies, such as cohort studies, and observational studies without a control group, is possible, since the instrument distinguishes between factors affecting the necessity and feasibility of a control group, and those affecting the necessity and feasibility of randomisation and blinding. Also, research with a historic control group may still be possible when, for some reason, randomisation proved unfeasible.

This study has some limitations. First, in the context of decision-making, the fundamental question is 'what to decide' given the particular context of an intervention including the existing evidence. The FIT instrument does not address this question, but addresses a vital question that needs to be answered beforehand, that is, the FIT instrument examines which study characteristics of a blinded RCT are necessary, and then, whether these characteristics are feasible. Thereafter, the appropriate evidence of effectiveness can be compared with the actual available evidence. If the available evidence lacks one or more of the study characteristics that were deemed necessary, the decision-maker may advise conditional reimbursement, the condition being that further evidence has to be assembled after reimbursement. However, the decision-maker may decide *not* to advise conditional reimbursement if the explanations for the existing lack of necessary study characteristics provided by the FIT instrument are considered legitimate. In this latter situation, the decision-maker will assume that no better-fitting evidence will probably appear in the future, and that a decision based on the available evidence will have to be made. Next to reasons for a suboptimal evidence base, additional factors are considered important in reimbursement decisions, such as disease severity and budget impact. These latter criteria concern equity in a given society as well as overall capacity for reimbursement, and fell beyond the scope of our research.

The second limitation concerns the instrument's completeness. Various disciplines have been active in this research area, making a complete overview complicated. Where clinical equipoise originates from ethics, the size of the patient population is an epidemiological argument to consider suboptimal evidence. Clinical

explanations, such as the urgency of an intervention, are also included in the instrument. Furthermore, there are practical reasons to deviate from randomisation, a control group or blinding. Therefore, we do not claim to have developed a definitive list of factors to assess the necessity and feasibility of randomisation, a control group and blinding. Nevertheless, we have tested the instrument on three consecutive reimbursement reports and found, besides its usefulness, that these reports did not identify any additional arguments. However, we recommend testing the instrument in future reimbursement decisions and adding any new argument wherever necessary.

Third, not only does the instrument's completeness require further testing, but so does its impact on the decision-making process in terms of efficiency, reliability (such as inter/intraobserver reliability) and additional forms of validity. The current validation phase was too short to properly address these issues. However, regardless of its impact on decisions, the instrument provides a degree of transparency in defining the appropriate evidence for an intervention, and this is in itself an important improvement over the status quo. The FIT instrument is frequently used by ZIN decision-makers since it was made available, which will lead to further modifications. ZIN intends to examine the FIT instrument's validity using the results of its application, and also intends to compare the results of its application to the available evidence for the interventions they have assessed using the FIT instrument.

Fourth, the FIT instrument focuses on the study characteristics of a blinded RCT, while an RCT has its disadvantages, such as limited generalisability and limited follow-up. Whereas efficacy might be more important for market authorisation to address causality, effectiveness might be more important from a decision-making perspective, which is the focus of this study. The optimal study design depends on the outcome of interest. If the outcome of interest appears to be a safety outcome, a blinded RCT may no longer be the optimal study design; the FIT instrument will show that randomisation is hard to achieve since 'outcomes occur in the distant future'. Therefore, the FIT instrument should be used for all outcomes of interest. Moreover, study design is not the only criterion to assess the quality of available evidence. This was pointed out by Guyatt *et al*,[7] who developed the GRADE system, which first qualifies evidence derived from RCTs as high quality evidence while observational studies are qualified as low quality evidence. Confidence in the estimate of effect is subsequently graded downwards and upwards based on criteria such as study limitations and inconsistency of results. For example, observational studies can be graded upwards if there is evidence of a dose–response relationship. While the GRADE approach focuses on assessing the quality of available evidence, we identified factors that outline which types of evidence can, in principle, be available, thereby providing arguments for reimbursement decisions when the available evidence does not match the evidence that is considered necessary and feasible.

Fifth, findings of this study were partly based on Dutch reimbursement reports, whereas reimbursement decisions vary across countries, since they are influenced by context, agency process, ability to engage in price negotiation and, perhaps, differences in social values.[37] Since most items in the FIT instrument are based on general epidemiological principles, the FIT instrument might be useful for reimbursement agencies in other countries as well.

Decision-makers strive to attain the most optimal evidence in the assessment of effectiveness, which mostly is a blinded RCT. There are, however, situations where an RCT is not specifically needed, for example, in case of a 'dramatic' or 'immediate' effect. Furthermore, if an RCT is needed, as a consequence of ethical as well as real world practical barriers, combined with the desire to efficiently allocate research funds, the most optimal evidence might still not be an RCT. Policy regarding the necessity and feasibility of different types of evidence of effectiveness would benefit from systematic guidance. Although the instrument needs further refinement, and although a critical appraisal of the factors influencing the necessity and feasibility of blinded RCTs and its constituent study characteristics remains essential, this instrument has the potential to support transparent, reproducible and well-founded decisions on appropriate evidence of effectiveness in medical specialist care.

Acknowledgements This work was financially supported by the Dutch National Health Care Institute. The authors would like to thank Bart Boksteijn, Joke Derksen, Angèl Link and Paula Staal for their cooperation in developing the FIT instrument, as well as to thank the four staff members who participated in the prospective validation of the instrument. Furthermore, the authors would like to thank the interviewees who gave their thoughts on the complexities of suboptimal evidence in reimbursement decisions. The authors are also grateful to Bert Boer for his critical comments on preliminary versions of this manuscript.

Contributors SdG collected the data (reimbursement reports, literature review and interviews), analysed the data, and drafted and revised the paper. She is guarantor. AR, JH and KR monitored data collection. All the authors helped in interpreting the findings, and writing and revising the draft paper.

Funding Dutch National Health Care Institute.

Competing interests SdG, AR, MV and KR received financial support from the Dutch National Health Care Institute for the submitted work.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1. Dutch National Health Care Institute (ZIN) (formerly named CVZ). *Assessment current medical science and practice [in Dutch: Beoordeling stand van de wetenschap en praktijk]*. Diemen, 2007. Report No: 27071300.

2. Sackett DL, Rosenberg WM, Gray JA, *et al*. Evidence based medicine: what it is and what it isn't. *BMJ* 1996;312:71–2.
3. Dutch National Health Care Institute (ZIN). *Assessment current medical science and practice (updated version 2015) [in Dutch: Beoordeling stand van de wetenschap en praktijk (geactualiseerde versie 2015)]*. Diemen, 2015. Report No: 2014116583.
4. Fischer KE. A systematic review of coverage decision-making on health technologies-evidence from the real world. *Health Policy* 2012;107:218–30.
5. Black N. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996;312:1215–18.
6. Rawlins M. De testimonio: on the evidence for decisions about the use of therapeutic interventions. *Lancet* 2008;372:2152–61.
7. Guyatt GH, Oxman AD, Vist GE, *et al*. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ* 2008;336:924–6.
8. Heymans JM, Kleijnen S, Verstijnen IM. 'Fitting' evidence preferable when evaluating effectiveness of interventions. *Ned Tijdschr Geneeskd* 2013;157:A5479.
9. Bowling A. Data collection methods in quantitative research: questionnaires, interviews and their response rates. In: Bowling A, ed. *Research methods in health, investigating health and health services*. Maidenhead, UK: Open University Press, 2002:257–72.
10. Dutch National Health Care Institute (ZIN) (formerly named CVZ). *Report on Metal on metal (MoM) resurfacing arthroplasty of the hip [in Dutch: MoM-heupprothese]*. Diemen, 2007. Report No: 27024808 (27052545).
11. Dutch National Health Care Institute (ZIN) (formerly named CVZ). *Report on preimplantation genetic diagnosis (PGD) with HLA typing for stem cell transplantation [in Dutch: Preïmplantatie genetische diagnostiek (PGD) in combinatie met HLA-typering van IVF-embryo's ten behoeve van eventuele stamceltransplantatie]*. Diemen, 2007. Report No: 27028502 (27070951).
12. Dutch National Health Care Institute (ZIN) (formerly named CVZ). *Assessment Diagnosis Treatment Combination 'Treatment for benign prostatic hyperplasia with KTP laser' [in Dutch: Beoordeling DBC 'de behandeling van benigne prostaat hyperplasie met KTP laser']*. Diemen, 2009. Report No: 28106169.
13. Dutch National Health Care Institute (ZIN) (formerly named CVZ). *Assessment Diagnosis Treatment Combination 'Transcatheter aortic valve implantation and Transcatheter pulmonary valve implantation' [in Dutch: Beoordeling DBC 'transkatheter aortaklep- en pulmonalisklepimplantatie']*. Diemen, 2009. Report No: 29022345.
14. Dutch National Health Care Institute (ZIN) (formerly named CVZ). *Report on indications for proton therapy (part 1): Intraocular tumours, chordomas/chondrosarcomas, paediatric tumours [in Dutch: Indicaties voor protonentherapie (deel 1): Intra-oculaire tumoren, chordomen/ chondrosarcomen, pediatrische tumoren]*. Diemen, 2010. Report No: 2010001046.
15. Dutch National Health Care Institute (ZIN) (formerly named CVZ). *Report on psychoanalysis and long-term psychoanalytic psychotherapy [in Dutch: Psychoanalyse en langdurige psychoanalytische psychotherapie]*. Diemen, 2010. Report No: 2010036278.
16. Dutch National Health Care Institute (ZIN) (formerly named CVZ). *Background report on the assessment current medical science and practice: Breast augmentation with autologous lipofilling [in Dutch: Achtergrondrapportage beoordeling stand van de wetenschap en praktijk: Mamma-augmentatie met autologe lipofilling]*. Diemen, 2008. Report No: 28048276.
17. Manchikanti L, Hirsch JA, Smith HS. Evidence-based medicine, systematic reviews, and guidelines in interventional pain management: Part 2: randomized controlled trials. *Pain Physician* 2008;11:717–73.
18. Bekkering GE, Kleijnen J. Procedures and methods of benefit assessments for medicines in Germany. *Eur J Health Econ* 2008;9 (Suppl 1):5–29.
19. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *BMJ* 1998;317:1185–90.
20. Devereaux PJ, Yusuf S. The evolution of the randomized controlled trial and its role in evidence-based decision making. *J Intern Med* 2003;254:105–13.
21. Kirtane AJ, Gupta A, Iyengar S, *et al*. Safety and efficacy of drug-eluting and bare metal stents: comprehensive meta-analysis of randomized trials and observational studies. *Circulation* 2009;119:3198–206.
22. Buchbinder R, Osborne RH, Ebeling PR, *et al*. A randomized trial of vertebroplasty for painful osteoporotic vertebral fractures. *N Engl J Med* 2009;361:557–68.
23. Higgins JPT, Altman DG, Sterne JAC (eds). Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Green S (editors). Cochrane Handbook for Systematic Reviews of Interventions. Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. www.cochrane-handbook.org
24. Hill AB. The environment and disease: association or causation? *Proc R Soc Med* 1965;58:295–300.
25. McCulloch P, Taylor I, Sasako M, *et al*. Randomised trials in surgery: problems and possible solutions. *BMJ* 2002;324:1448–51.
26. Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med* 1987;317:141–5.
27. Sege RD, De Vos E. Evidence-based health care for children: what are we missing? *Issue Brief (Commonw Fund)* 2010;85:1–14.
28. Dutch National Health Care Institute (ZIN) (formerly named CVZ). *Background report on the assessment current medical science and practice: Sacral neurostimulation/ neuromodulation for fecal incontinence [in Dutch: Achtergrondrapportage beoordeling stand van de wetenschap en praktijk: Sacrale neurostimulatie/ neuromodulatie bij faecale incontinentie]*. Diemen, 2008. Report No: 28085299.
29. Khoury MJ, Coates RJ, Evans JP. Evidence-based classification of recommendations on use of genomic tests in clinical practice: dealing with insufficient evidence. *Genet Med* 2010;12:680–3.
30. Ergina PL, Cook JA, Blazeby JM, *et al*. Challenges in evaluating surgical innovation. *Lancet* 2009;374:1097–104.
31. Craig P, Dieppe P, Macintyre S, *et al*. Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ* 2008;337:a1655.
32. Campbell M, Fitzpatrick R, Haines A, *et al*. Framework for design and evaluation of complex interventions to improve health. *BMJ* 2000;321:694–6.
33. Dutch National Health Care Institute (ZIN) (formerly named CVZ). *Report on rheopheresis in the treatment of dry macular degeneration [in Dutch: Rheopherese-behandeling bij droge maculadegeneratie]*. Diemen, 2010. Report No.: 2010081059 (2010122609).
34. Guyatt GH, Oxman AD, Kunz R, *et al*. What is "quality of evidence" and why is it important to clinicians? *BMJ* 2008;336:995–8.
35. Dutch National Health Care Institute (ZIN) (formerly named CVZ). *Report on vertebroplasty en ballon kyphoplasty [in Dutch: Vertebroplastiek en ballonkyphoplastiek]*. Diemen, 2010. Report No.: 29060581 (2010106817).
36. van Loon J, Grutters J, Macbeth F. Evaluation of novel radiotherapy technologies: what evidence is needed to assess their clinical and cost effectiveness, and how should we get it? *Lancet Oncol* 2012;13:e169–77.
37. Clement FM, Harris A, Li JJ, *et al*. Using effectiveness and cost-effectiveness to make drug coverage decisions: a comparison of Britain, Australia, and Canada. *JAMA* 2009;302:1437–43.